

Visual In-Context Learning for Few-Shot Eczema Segmentation

Neelesh Kumar, Oya Aran, Venugopal Vasudevan

Procter and Gamble, Mason, OH 45040 USA
`{kumar.n.40, aran.o, venu}@pg.com`

Abstract. Automated diagnosis of eczema from digital camera images is crucial for developing applications that allow patients to self-monitor their recovery. An important component of this is the segmentation of eczema region from such images. Current methods for eczema segmentation rely on deep neural networks such as convolutional (CNN)-based U-Net or transformer-based Swin U-Net. While effective, these methods require high volume of annotated data, which can be difficult to obtain. Here, we investigate the capabilities of visual in-context learning that can perform few-shot eczema segmentation with just a handful of examples and without any need for retraining models. Specifically, we propose a strategy for applying in-context learning for eczema segmentation with a generalist vision model called SegGPT. When benchmarked on a dataset of annotated eczema images, we show that SegGPT with just 2 representative example images from the training dataset performs better (mIoU: 36.69) than a CNN U-Net trained on 428 images (mIoU: 32.60). We also discover that using more number of examples for SegGPT may in fact be harmful to its performance. Our result highlights the importance of visual in-context learning in developing faster and better solutions to skin imaging tasks. Our result also paves the way for developing inclusive solutions that can cater to minorities in the demographics who are typically heavily under-represented in the training data.

Keywords: Eczema · Segmentation · In-context · Vision.

1 Introduction

Eczema is one of the most common skin disorders, with over 10% of the population affected by it in the United States alone [2]. While there can be no substitute for expert dermatologists, automated diagnosis of the disease can enable patients to self-monitor their recovery using self-acquired images, potentially leading to more effective recovery due to the psychological effects [14,10]. An important part of this automated analysis is developing a sufficiently robust algorithm that can accurately segment the eczema region from the patient-acquired digital camera images. [15,8,23].

The state-of-the-art approaches for automated eczema segmentation rely on deep neural networks (DNN) for supervised learning, with the most pop-

ular architecture being a convolutional neural network-based U-Net (CNN U-Net) [18,5]. More recent approaches, such as Swin U-Net, leverage attention-based transformers that can capture long-range dependencies [7]. These methods have demonstrated remarkable performance improvements over traditional approaches that relied on hand-crafted visual features [8,24]. However, a common theme across all these methods is the need for sufficient data to train the DNN, which increases with the increasing complexity of the network architecture [3].

For applications such as skin lesion segmentation that require expert annotations, acquiring a large dataset can be prohibitive from the point of view of both cost and time [26]. This is further exacerbated by the need for a diverse dataset covering all sources of variations such as skin tone, images from different body parts, and varying levels of disease [25,15]. Current methods to train high-capacity networks using limited data employ transfer learning techniques such as domain adaptation, knowledge distillation, or finetuning [27,11]. While these techniques have proven to be effective for many learning tasks, including skin segmentation [4], these methods still require enough labeled and unlabeled training data to work well.

A long-standing goal in artificial intelligence is to learn task-agnostic representations that can be used across various tasks without requiring further learning [12]. This few/zero-shot form of learning eliminates the necessity of training task-specific models and hence any need for training data. While the goal seems ambitious, we are already seeing an outpour of results from the Natural Language Processing (NLP) community [21,17]. Known popularly as large language models (LLMs), these models learn task-agnostic representations through pre-training on a large corpus of text data, and exhibit remarkable generalization on NLP downstream applications[17]. The operating principle is In-context Learning, where a few domain-specific input-output pairs are provided as in-context examples (prompts) to the model, along with the input test example [32,20]. The prompt and the input test example are used to predict the corresponding output for the test example without updating model weights [32].

A growing body of work is now attempting to replicate this success of NLP for vision tasks using a paradigm known as visual in-context learning [6,32,28,29,19,16]. The idea is similar to that in NLP: a high-capacity network such as a Vision Transformer (ViT) [9] is trained on a large corpus of images using Masked Image Modeling (MIM) [31,13]. The complexity of the MIM task, i.e., predicting large missing patches in the image, forces the network to gain semantic and contextual understanding of the images resulting in it learning powerful general representations [13]. These learned representations are versatile enough to perform various downstream vision tasks such as keypoint detection, segmentation, depth estimation, etc., following prompt-style input representation [28,29]. With visual in-context learning, it is possible to perform few-shot segmentation of eczema images without requiring an abundance of training data.

This work presents an automated approach to few-shot eczema segmentation using visual in-context learning (Figure 1). Specifically, we employ a pretrained generalist vision model called SegGPT [29] and evaluate it for eczema segmen-

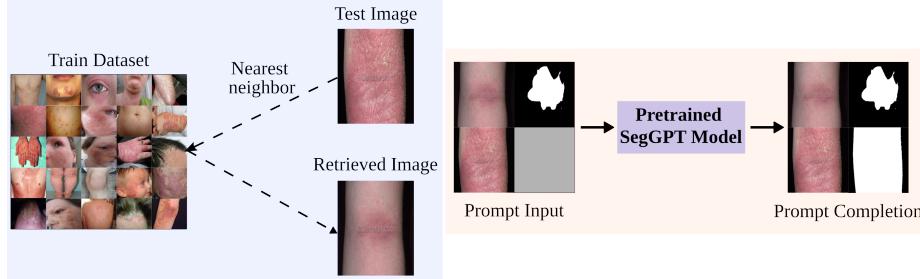


Fig. 1. Overview. For each test image, its k -nearest neighbors are retrieved from the training dataset. The retrieved images and their masks, along with the test image are stitched together to construct a prompt that is fed to a pretrained SegGPT model. The task of the model is to predict the missing output image for the given test image.

tation on a dataset of skin images acquired from the web and a consumer study. We report that SegGPT with just two example images in the prompt performs better (mIoU: 36.69) than a CNN U-Net trained on 428 images (mIoU: 32.60). The key to this result is the strategy for prompt selection: the examples in the prompt must be representative of the task. Our simple framework for prompt selection retrieves nearest neighbors of the test image from the training dataset, and uses them as examples to construct the prompt. We also discover that the performance of SegGPT is strongly dependent on the number of examples in the prompt, but surprisingly does not have a linear relation. Rather, the performance increases up to a certain number $k < 4$, and then starts decreasing. Our results highlight the promise of foundational generalist vision models in developing faster and better solutions for skin lesion segmentation tasks, paving the way for developing inclusive solutions that can also cater to minorities in the demographics who are typically heavily under-represented in the training data.

2 Methods

The core of our method is a visual in-context learning model called SegGPT [29]. The input to the model is a visual prompt which is constructed by stitching together example input-output image pairs. The test input image with a blank output image is appended to the prompt. The task of the model is to complete the prompt, i.e., predict the blank output image for the given test input image (Figure 1) without updating any model weights.

2.1 SegGPT Training

The segmentation task in SegGPT is formulated as an image in-painting task [28]: given an input image, the prediction task is to inpaint the desired but missing output image. This allows for a standard image input and image output interface for the model, because of which the model can be trained on a large

corpus of vision data irrespective of the actual vision task. So long as the input and output of the vision task can be represented as images, the model can leverage the dataset associated with the task for training.

The training of SegGPT is based on MIM [13]. During training, two images from the same segmentation task are stitched into a larger image, along with their corresponding masks. MIM is then applied to the pixels of output mask images. The large masking ratio forces the model to gain contextual and semantic understanding of the image to complete the in-painting task [13,31]. This understanding allows the trained model to understand what areas to segment from the example input-output pairs in the prompt.

The model in SegGPT employs a vanilla vision transformer (ViT) [9] as the encoder consisting of stacked transformer blocks. A three layer head comprising of a linear layer, a 3×3 convolution layer, and another linear layer is used to map the features of each patch to the original resolution. A simple smooth l_1 regression loss is used to train the network.

2.2 Prompt Selection

The performance of visual in-context learning models depends strongly upon the choice of prompts [32]. Inspired from the results in [32], we adopt a similarity-based method to retrieve the prompt from the training dataset. Specifically, for each test image, we compute its k nearest neighbors from the training dataset based on a distance metric such as Frobenius norm and structural similarity index (SSIM). These neighbors are then used to construct the prompts for the given test image.

3 Experiments and Results

The goals of our experiments were to investigate whether visual in-context learning can perform as well as traditional methods that rely on training on large datasets. To that effect, we compared the performance of a pretrained SegGPT against a CNN-based U-Net trained to segment eczema images.

3.1 Dataset description

Our dataset consists of 528 high-resolution images collected from two primary sources: a) public Dermnet dataset of eczema images [1], and b) an in-house consumer study where images were taken directly by the participants using their smartphone cameras. The resulting dataset has multiple sources of variations: eczema images from different body parts, varying skin tones, and varying levels of eczema along with varying illumination, background, etc. which increases the complexity of the segmentation task.

The dataset was labelled using human annotators. The human annotators were provided written instructions from domain experts on how to create the masks. Several examples were shown to them before they started the task. Each

image was annotated by one human annotator. The resulting masks were then inspected by multiple domain experts and were found to be satisfactory.

All images were resized to 448 x 448 which is the input dimension for the pretrained SegGPT model. Further, the images were normalized using z-score normalization using the ImageNet dataset statistics. The dataset was partitioned into 428 training images and 100 evaluation images. The baseline U-Net model used all 428 images for its training. The SegGPT model used a handful of k samples from the training set to construct prompts.

3.2 Baseline for comparison

As baseline, we used CNN-based U-Net which is widely used for skin segmentation [22]. The architecture was a typical 5-stage U-Net comprising of serial contracting and expansive paths [22]. Similar to [22], the contracting path consisted of repeated two 3x3 unpadded convolutions, followed by batch normalization, ReLU and maxpooling with window size 2x2. The number of channels were doubled after every stage. The expansive path followed the inverse operations of the contracting path- upsampling of feature maps followed by repeated up-convolutions that halve the number of channels, with ReLU non-linearity.

The network was trained on 428 training images using Adam optimizer with a small weight decay factor for additional regularization. Learning rate was set to $1e - 4$. Batch size was set to 8. The network was trained for 50 epochs with cross-entropy loss function.

To evaluate, we use the intersection over union score (mIoU) averaged over all test images. As the name suggest, mIoU measures the area of overlap between the predicted segmentation mask and the ground-truth mask, and hence is a more suitable metric to measure segmentation performance than pixel-wise accuracy.

3.3 SegGPT Network Architecture and Prompt Retrieval

We used an off-the-shelf pretrained SegGPT model that employs a vanilla ViT-large as its encoder. The model was trained to optimize the smooth l_1 regression loss. The authors in [29] pretrained the model on a large collection of benchmark segmentation datasets: ADE20K, COCO panoptic, Cityscapes, COCO semantic, LIP person, PASCAL VOC, PACO, iSAID and loveDA aerial, CHASEDB, DRIVE, HRF and STARE retinal vessel.

For prompt selection, we selected k nearest neighbor of each test image from the training dataset, varying k from 1 to 15 (Figure 1). We employed the following distance metrics: i) the commonly used Frobenius norm to compute euclidean difference between two matrices; ii) SSIM, which is a better indicator of similarity between images by virtue of being a perception-based metric.

3.4 Segmentation of Eczema

We evaluated pretrained SegGPT and baseline U-Net for segmentation of eczema. While U-Net achieved an mIoU score of 32.60 on the test images, the pretrained

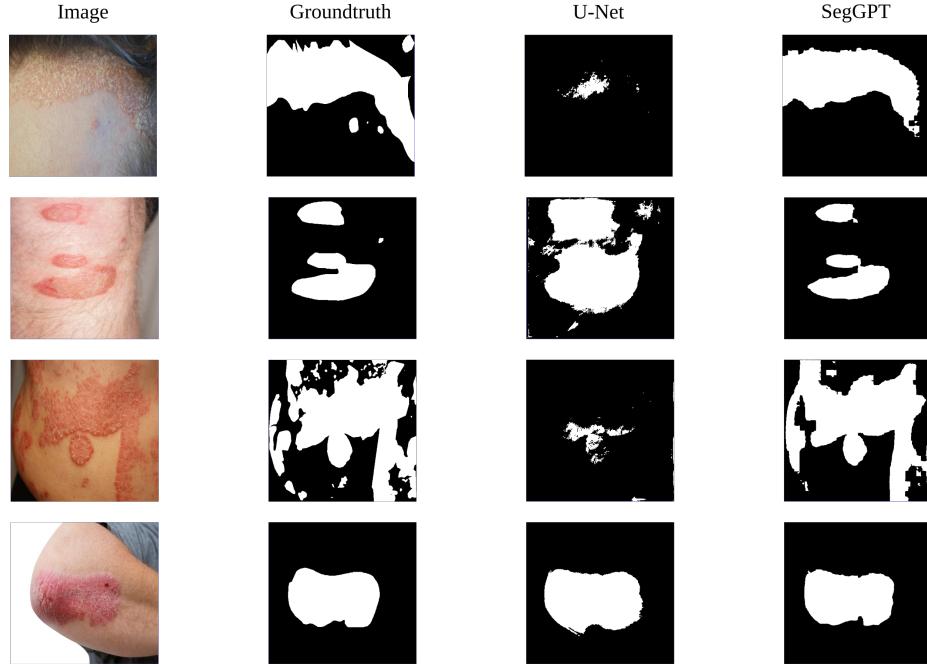


Fig. 2. Qualitative comparison of eczema segmentation performance of U-Net and SegGPT for 4 example images. SegGPT produces masks that are closer to the groundtruth.

Table 1. Segmentation Performance

Method	mIoU
U-Net	32.60
SegGPT ($k = 2$; SSIM)	36.69

SegGPT with the optimal hyperparameter $k = 2$ achieved an mIoU score of 36.69 (Table 1). Not only did SegGPT outperform U-Net, it did so by using only 2 representative images from the training set and without changing any of its weights. This amounts to a 12.6% increase in performance with a 213 times decrease in the training data requirement. The key is that the images in the prompt must be representative of the segmentation task. The quantitative improvement is further highlighted in the qualitative comparison in Figure 2, where SegGPT produced comparable or higher quality masks than U-Net.

Effect of k and distance metric To understand the dependence of the performance of SegGPT on the number of examples used to construct the prompt, we evaluated its performance for a range of values of k . We show in Figure 3 that the mIoU increases as we increase k up to a certain level, and then starts to decrease. This indicates that picking higher number of examples to construct the

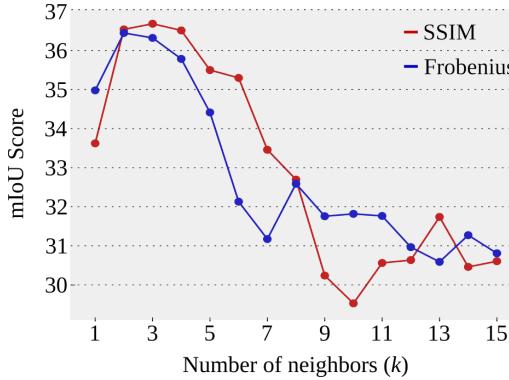


Fig. 3. Dependence of the segmentation performance on the number of neighbors, k and the distance metric. The performance increases up until a low value of k and then decreases. Using SSIM as distance metric results in marginally better performance than using Frobenius norm.

prompt may not result in better performance. Although this may seem counter-intuitive, we believe that the reason for this observation is as follows: As we increase k , we pick examples that are further away from the test image. Given the limited size of the dataset, images that are too far from the test image may not be representative. To highlight this point, we show the k nearest neighbors for two representative test images for both the distance metric in Figure 4. Since the SegGPT model is not retrained/conditioned on the prompt, it is not capable of distinguishing examples that may not be representative of the segmentation task, and instead places equal emphasis on all the examples in the prompt. As one might expect, the performance obtained when SSIM is used as a distance metric is slightly better than when Frobenius norm is used (Figure 3).

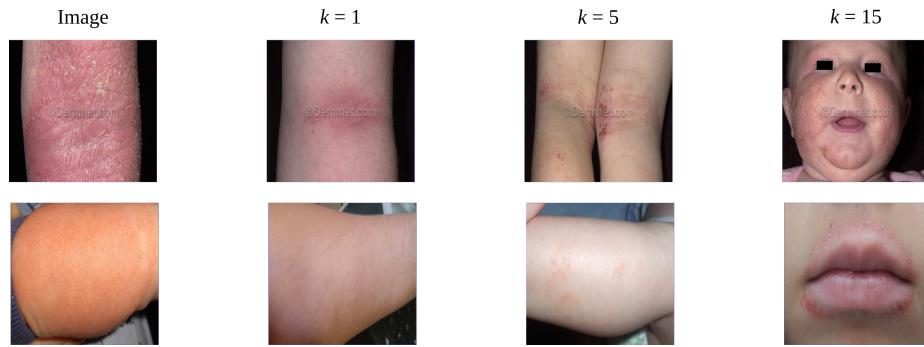


Fig. 4. Nearest neighbor images for two example test images. As k increases, the retrieved neighbor is further away from the test image.

4 Discussion

In this work, we present a visual in-context learning approach for segmentation of eczema from patient-acquired images. We showed that with just two examples of the segmentation task in the prompt, and the right prompt-selection strategy, the pretrained SegGPT can perform better than the state-of-the-art CNN-based U-Net despite the fact that the latter sees 428 examples in its training. Our result adds to the mounting evidence that learning task-agnostic features on large diverse datasets with high-capacity models eliminates the need for performing any training or finetuning for downstream tasks [17,20,19,16].

A diverse dataset for medical and skin imaging that is representative of the underlying demographics is hard to get and even harder to annotate [26]. The current methods to deal with the limited data use either self-supervision which relies on having enough unlabelled data, or finetuning which again requires sufficient labelled data [27]. On the other hand, visual in-context learning can leverage a handful of representative labeled examples to perform the task with competitive performance, drastically reducing the time and effort for data collection. While our current method uses the entire training dataset to search for representative examples, in a consumer-facing application, patients will need to provide just 1 or 2 annotated images for the model to make accurate predictions.

In addition, the approach holds significance for application areas where it is impractical to wait for enough labelled data to arrive before accurate predictions can be made. Such application areas include consumer-facing applications where patients can self-monitor the trajectory of the improvement of their skin condition when following a treatment protocol [14,10,30]. In such cases, asking patients to wait until we have enough data from them may not be prudent, and in-context learning can play a key role.

Given the fact that the performance of in-context learning depends strongly on the choice of prompts, more systematic ways of prompt retrieval can be investigated. The current approach relies on pixel-level distance between the two images. However, measuring similarity between images at feature-level may result in more effective prompts. The work in [32] presents two such approaches relying on supervised and unsupervised learning. Additionally, although the focus of this work is to obtain competitive performance without any training or finetuning, if additional performance improvement is desired, the ViT can further be finetuned on domain-specific data.

A key issue in skin imaging is the under-representation of certain demographics in the training data, as a result of which methods may be biased towards heavily represented groups. In-context learning, with its ability to generalize from just a few data points, has the potential to tackle this issue. More experiments are needed on benchmark evaluation datasets that contain data from under-represented groups to confirm the hypothesis. Overall, our work highlights the importance of the increasingly-popular in-context learning framework, and the possible directional shift from the traditional train-test-finetune paradigm.

References

1. Dermnet dataset, <https://www.dermnetnz.org/>
2. Abuabara, K., Magyari, A., McCulloch, C.E., Linos, E., Margolis, D.J., Langan, S.M.: Prevalence of atopic eczema among patients seen in primary care: Data from the health improvement network. *Annals of internal medicine* **170**(5) (2019)
3. Adadi, A.: A survey on data-efficient algorithms in big data era. *Journal of Big Data* **8**(1), 24 (2021)
4. Al Mamun, M., Uddin, M.S.: A comparative study among segmentation techniques for skin disease detection systems. In: Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020. pp. 155–167. Springer (2020)
5. Anand, V., Gupta, S., Koundal, D., Singh, K.: Fusion of u-net and cnn model for segmentation and classification of skin lesion from dermoscopy images. *Expert Systems with Applications* **213**, 119230 (2023)
6. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting. *Advances in Neural Information Processing Systems* **35**, 25005–25017 (2022)
7. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swinunet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
8. Ch'ng, Y.K., Nisar, H., Yap, V.V., Yeap, K.H., Tang, J.J.: Segmentation and grading of eczema skin lesions. In: 2014 8th International Conference on Signal Processing and Communication Systems (ICSPCS). pp. 1–5. IEEE (2014)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Ehlers, A., Stangier, U., Gieler, U.: Treatment of atopic dermatitis: a comparison of psychological and dermatological approaches to relapse prevention. *Journal of consulting and clinical psychology* **63**(4), 624 (1995)
11. Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R.: A brief review of domain adaptation. *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020* pp. 877–894 (2021)
12. Goertzel, B.: Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence* **5**(1), 1 (2014)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
14. Horne, D.J.d.L., White, A.E., Varigos, G.A.: A preliminary study of psychological therapy in the management of atopic eczema. *British Journal of Medical Psychology* **62**(3), 241–248 (1989)
15. Hurault, G., Pan, K., Mokhtari, R., Olabi, B., Earp, E., Steele, L., Williams, H.C., Tanaka, R.J.: Detecting eczema areas in digital images: an impossible task? *JID Innovations* **2**(5), 100133 (2022)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
17. Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., et al.: Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852 (2023)

18. Nisar, H., Tan, Y.R., Yeap, K.H.: Eczema skin lesions segmentation using deep neural network (u-net). In: Advances in Non-Invasive Biomedical Signal Sensing and Processing with Machine Learning, pp. 229–262. Springer (2023)
19. Oquab, M., Dariseti, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
21. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
23. Roy, K., Chaudhuri, S.S., Ghosh, S., Dutta, S.K., Chakraborty, P., Sarkar, R.: Skin disease detection based on different segmentation techniques. In: 2019 international conference on opto-electronics and applied optics (Optronix). pp. 1–5. IEEE (2019)
24. Schnürle, S., Pouly, M., vor der Brück, T., Navarini, A., Koller, T.: On using support vector machines for the detection and quantification of hand eczema. In: ICAART (2). pp. 75–84 (2017)
25. Seth, D., Cheldize, K., Brown, D., Freeman, E.E.: Global burden of skin disease: inequities and innovations. Current dermatology reports **6**, 204–210 (2017)
26. Shi, X., Dou, Q., Xue, C., Qin, J., Chen, H., Heng, P.A.: An active learning approach for reducing annotation cost in skin lesion analysis. In: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10. pp. 628–636. Springer (2019)
27. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: Artificial Neural Networks and Machine Learning—ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, 2018, Proceedings, Part III 27. pp. 270–279. Springer (2018)
28. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6830–6839 (2023)
29. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Segmenting everything in context. arXiv preprint arXiv:2304.03284 (2023)
30. Wittkowski, A., Richards, H.L.: How beneficial is cognitive behaviour therapy in the treatment of atopic dermatitis? a single-case study. Psychology, health & medicine **12**(4), 445–449 (2007)
31. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)
32. Zhang, Y., Zhou, K., Liu, Z.: What makes good examples for visual in-context learning? arXiv preprint arXiv:2301.13670 (2023)